

Recursive partitioning and multi-scale modeling on conditional densities

Li Ma *

*Department of Statistical Science
Duke University
Durham, NC 27708-0251, USA
e-mail: li.ma@duke.edu*

Abstract: We introduce a nonparametric prior on the conditional distribution of a (univariate or multivariate) response given a set of predictors. The prior is constructed in the form of a two-stage generative procedure, which in the first stage recursively partitions the predictor space, and then in the second stage generates the conditional distribution by a multi-scale nonparametric density model on each predictor partition block generated in the first stage. This design allows adaptive smoothing on both the predictor space and the response space, and it results in the full posterior conjugacy of the model, allowing exact Bayesian inference to be completed analytically through a forward-backward recursive algorithm without the need of MCMC, and thus enjoying high computational efficiency (scaling linearly with the sample size). We show that this prior enjoys desirable theoretical properties such as full L_1 support and posterior consistency. We illustrate how to apply the model to a variety of inference problems such as conditional density estimation as well as hypothesis testing and model selection in a manner similar to applying a parametric conjugate prior, while attaining full nonparametricity. A real data example from flow cytometry is also provided in which we also illustrate the substantial computational efficiency of the method.

AMS 2000 subject classifications: Primary 62F15, 62G99; secondary 62G07.

Keywords and phrases: Pólya tree, multi-resolution inference, Bayesian nonparametrics, density regression.

1. Introduction

In recent years there has been growing interest in nonparametrically modeling probability densities based on multi-scale partitioning of the sample space. A prime example in the Bayesian nonparametric literature is the Pólya tree (PT) [9, 18, 25] and its extensions [14, 15, 40, 17, 21]. In particular, Wong and Ma [40] introduced randomization into the partitioning component (involving both random selection of partition directions as well as optional stopping) of the PT framework, which enhances the model's ability to approximate the shape and smoothness of the underlying density. A PT model with these features is called an optional Pólya tree (OPT).

A further desirable feature of the OPT is the computational ease for carrying out inference. In turns out that the extra component of randomized partitioning does not impair the conjugacy enjoyed by the PT—after observing i.i.d. data, the corresponding posterior of an OPT is still an OPT, that is, the same generative procedure for random probability distributions with its parameters updated to their posterior values. Moreover, the corresponding posterior parameter values can be computed *exactly* through a sequence of recursive computations, which is in essence a forward-backward algorithm [20] as pointed out in [21]. This, together with the constructive nature of the prior, allows one to draw samples from the exact posterior directly without resorting to Markov Chain Monte Carlo (MCMC) procedures, and to compute various summary statistics of the posterior analytically. Furthermore, the marginal posterior of the random partitioning adapts to the underlying structure of the data—the sample space will with high posterior probability be more finely divided in places where the underlying distribution has richer structure, i.e. less uniform topological shape.

Motivated by the computational efficiency and theoretical properties of the OPT, which is tied to its use of recursive random partitioning, we aim to further exploit the random recursive partitioning idea in the context of multi-scale density modeling, and build such a model for *conditional* densities for a response (vector) \mathbf{Y} given a predictor (vector) \mathbf{X} .

*Supported in part by NSF grants DMS-1309057 and DMS-1612889, and a Google Faculty Research Award.

Many inference tasks involve the estimation, prediction, and testing regarding conditional distributions, and nonparametric inference on conditional densities has been studied from both frequentist and Bayesian perspectives. Many frequentist works are based on kernel estimation methods [7, 13, 8], and they achieve proper smoothing through bandwidth selection, which often involves resampling procedures such as cross-validation [1, 16, 8] and the bootstrap [13]. In Bayesian nonparametrics, inference on conditional distributions is often referred to as covariate-dependent distribution modeling, and existing methods fall into two categories. The first is methods that construct priors for the joint distribution of the response and the predictors, and then use the induced conditional distribution for inference. Some examples are [26, 31, 27, 35], which propose using mixtures of multivariate normals as the model for joint distributions, along with different priors for the mixing distribution. The other category is methods that construct conditional distributions directly without specifying the marginal distribution of the predictors. Many of these methods are based on extending the stick breaking construction for the Dirichlet Process (DP) [33]. Some notable examples, among others, are proposed in [23, 4, 10, 12, 6, 3, 30]. Four recent works in this category do not utilize stick breaking. In [37], the authors propose to use the logistic Gaussian process [19, 36] together with subspace projection to construct smooth conditional distributions. In [17], the authors incorporate covariate dependency into tail-free processes by generating the conditional tail probabilities from covariate-dependent logistic Gaussian processes, and propose a mixture of such processes as a way for modeling conditional distributions. In [38] the authors introduce the covariate-dependent multivariate Beta process, and use it to generate the conditional tail probabilities of Pólya trees. More recently, in [34] the authors use the tensor product of B-splines to construct a prior for conditional densities, and incorporate a variable selection feature. Inference using these Bayesian nonparametric priors on conditional distributions generally relies on intense MCMC sampling.

We introduce a new prior, called the conditional optional Pólya tree, for the conditional density of \mathbf{Y} given \mathbf{X} , in the form of a two-stage generative procedure consisting of first randomly partitioning the *predictor* space $\Omega_{\mathbf{X}}$, and then for each predictor partition block, generates the response distribution on each block using an OPT, which implicitly employs a further random partitioning of the *response* space $\Omega_{\mathbf{Y}}$. We show that this new prior also enjoys the desirable theoretical properties of the OPT prior—namely large support, posterior consistency, and posterior conjugacy, and its posterior parameters can also be computed exactly through forward-backward recursion. Under this two-stage design, the posterior distribution on the partitions reflect the structure of the conditional distribution at two levels—first, the predictor space will be partitioned finely in parts where the conditional distribution changes most abruptly, shedding light on how the conditional distribution depends on the predictors and providing a means for model selection; second, the response space will be divided adaptively for different locations of the predictor space, to capture the local structure of the conditional density through the OPTs.

The rest of the paper is organized as follows. In [Section 2](#) we introduce our two-stage prior and show that it is fully nonparametric (with full L_1 support) for conditional densities. In addition, we make a connection to Bayesian CART and show that our method can be considered a nonparametric version of the latter. In [Section 3](#) we show the full conjugacy of the model, derive the exact form of the posterior through forward-backward recursion, and thereby provide a recipe for carrying out Bayesian inference using the prior. We also prove the posterior consistency of such inference. In [Section 4](#) we discuss practical computational issues in implementing the inference. In [Section 5](#) we provide four simulation examples to illustrate the work of our method. The first two are for estimating conditional densities, and the last two concern model selection and hypothesis testing. In [Section 6](#) we apply the proposed method to estimating conditional densities in a flow cytometry data set involving a large number (455,472) of observations, and demonstrate the computational efficiency of the method. [Section 7](#) concludes with some discussions. All proofs are given in the Appendix.

2. Conditional optional Pólya trees

In this section we introduce our proposed prior constructively in terms of a two-stage generative procedure that produces random conditional densities. First we introduce some notions and notations that will be used throughout. Let each observation be a predictor-response pair (\mathbf{X}, \mathbf{Y}) , where \mathbf{X} denotes the predictor (or covariate) vector and \mathbf{Y} the response (vector) with $\Omega_{\mathbf{X}}$ being the predictor space and $\Omega_{\mathbf{Y}}$ the response space. In this work we consider sample spaces that are either finite spaces, compact Euclidean rectangles, or a product of the two, and $\Omega_{\mathbf{X}}$ and $\Omega_{\mathbf{Y}}$ do not have to be of the same type. (See for instance [Example 3](#).) Let

$\mu_{\mathbf{X}}$ and $\mu_{\mathbf{Y}}$ be the “natural” measures on $\Omega_{\mathbf{X}}$ and $\Omega_{\mathbf{Y}}$. (That is, the counting measure for finite spaces, the Lebesgue measure for Euclidean rectangles, and the corresponding product measure if the space is a product of the two.) Let $\mu = \mu_{\mathbf{X}} \times \mu_{\mathbf{Y}}$ be the “natural” product measure on the joint sample space $\Omega_{\mathbf{X}} \times \Omega_{\mathbf{Y}}$.

A *partition rule* \mathcal{R} on a sample space Ω specifies a collection of possible ways to divide any subset A of Ω into a number of smaller sets. For example, for $\Omega = [0, 1]^k$, the unit rectangle in \mathbb{R}^k , the *coordinate-wise dyadic mid-split rule* allows each rectangular subset A of Ω whose sides are parallel to the k coordinates to be divided into two halves at the middle of the range of each coordinate. For simplicity, in this work we only consider partition rules that allow a *finite* number of ways for dividing each set. Such partition rules are said to be *finite*. (Interested readers can refer to [22, Sec. 2] for a more detailed treatment of partition rules and to Examples 1 and 2 in [40] for examples of the coordinate-wise dyadic mid-split rule over Euclidean rectangles and 2^k contingency tables.)

We are now ready to introduce our prior for conditional densities as a two-stage constructive procedure.

Stage I. Predictor partition: We randomly partition $\Omega_{\mathbf{X}}$ according to a given partition rule $\mathcal{R}_{\mathbf{X}}$ on $\Omega_{\mathbf{X}}$ in the following recursive manner. Starting from $A = \Omega_{\mathbf{X}}$, draw a Bernoulli variable

$$S(A) \sim \text{Bernoulli}(\rho(A)).$$

That is, $P(S(A) = 1) = \rho(A)$. If $S(A) = 1$, then the partitioning procedure on A terminates and we arrive at a trivial partition of a single block over A . (Thus $S(A)$ is called the stopping variable, and $\rho(A)$ the stopping probability.) If instead $S(A) = 0$, then we randomly select one out of the possible ways for dividing A under $\mathcal{R}_{\mathbf{X}}$ and partition A accordingly. More specifically, if there are $N(A)$ ways to divide A under $\mathcal{R}_{\mathbf{X}}$, we randomly draw

$$J(A) \in \{1, 2, \dots, N(A)\} \text{ such that } P(J(A) = j) = \lambda_j(A) \text{ for } j = 1, 2, \dots, N(A) \text{ with } \sum_{j=1}^{N(A)} \lambda_j(A) = 1$$

and partition A in the j th way if $J(A) = j$. (We call $\boldsymbol{\lambda}(A) = (\lambda_1(A), \lambda_2(A), \dots, \lambda_{N(A)}(A))$ the partition selection probabilities for A .) Let $K^j(A)$ be the number of child sets that arise from this partition, and let $A_1^j, A_2^j, \dots, A_{K^j(A)}^j$ denote these children. We then repeat the same partition procedure, starting from the drawing of a stopping variable, on each of these children.

The following lemma states that as long as the stopping probabilities are (uniformly) away from 0, this random recursive partitioning procedure will eventually terminate almost everywhere and produce a well-defined partition of $\Omega_{\mathbf{X}}$.

Lemma 1. *If there exists a $\delta > 0$ such that the stopping probability $\rho(A) > \delta$ for all $A \subset \Omega_{\mathbf{X}}$ that could arise after a finite number of levels of recursive partition, then with probability 1 the recursive partition procedure on $\Omega_{\mathbf{X}}$ will stop $\mu_{\mathbf{X}}$ a.e.*

Stage II. Generating conditional densities: Next we move onto the second stage of the procedure to generate the conditional density of the response \mathbf{Y} on each of the predictor partition blocks generated in Stage I. Specifically, for each stopped subset A on $\Omega_{\mathbf{X}}$ produced in Stage I, we let the conditional distribution of \mathbf{Y} given $\mathbf{X} = x$ be the same across all $x \in A$, and generate this (conditional) distribution on $\Omega_{\mathbf{Y}}$, denoted as $q_{\mathbf{Y}}^{0,A}$, from a “local” prior. When the response space $\Omega_{\mathbf{Y}}$ is a finite, $q_{\mathbf{Y}}^{0,A}$ is a multinomial distribution, and so a simple choice of such a local prior is the Dirichlet prior: $q_{\mathbf{Y}}^{0,A} \sim \text{Dirichlet}(\boldsymbol{\alpha}_{\mathbf{Y}}^A)$ where $\boldsymbol{\alpha}_{\mathbf{Y}}^A$ represents the pseudo-count hyperparameters of the Dirichlet. In this case, we note that the two-stage prior essentially reduces to a version of the Bayesian CART proposed by Chipman et al in [2] for the classification problem.

When $\Omega_{\mathbf{Y}}$ is infinite (or finite but with a large number of elements), one may restrict $q_{\mathbf{Y}}^{0,A}$ to be from a parametric family. For example, when $\Omega_{\mathbf{Y}} = \mathbb{R}$, one may require $q_{\mathbf{Y}}^{0,A}$ to be normal with some mean μ_A and variance σ_A^2 , and let $\mu_A | \sigma_A^2 \sim N(\mu_0, \sigma^2)$ and $\sigma_A^2 \sim \text{inverse-Gamma}(\nu/2, \nu\kappa/2)$. The two-stage prior again reduces to a Bayesian CART, this time for the regression problem [2].

The focus of our current work, however, is on the case when no parametric assumptions are placed on the conditional density. To this end, one can draw $q_{\mathbf{Y}}^{0,A}$ from a nonparametric prior. A desirable choice for the

local prior, which will result in analytic simplicity and computational efficiency as we will later show, is the optional Pólya tree (OPT) distribution [22, Sec. 2]:

$$q_Y^{0,A} \sim \text{OPT}(\mathcal{R}_Y^A; \rho_Y^A, \lambda_Y^A, \alpha_Y^A)$$

where \mathcal{R}_Y^A denotes a partition rule on Ω_Y and ρ_Y^A , λ_Y^A , and α_Y^A are the hyperparameters. Note that in general we allow the partition rule for these “local” OPTs to depend on A as indicated in the superscript, but adopting a common partition rule on Ω_Y —that is to let $\mathcal{R}_Y^A \equiv \mathcal{R}_Y$ for all A —will suffice for most problems. In the rest of the paper, unless stated otherwise we assume that a common rule \mathcal{R}_Y is adopted. This completes the description of our two-stage procedure. We are now ready to present a formal definition of the prior.

Definition 1. A conditional distribution that arises from the above two-stage procedure is said to have a *conditional optional Pólya tree* (cond-OPT) distribution. The hyperparameters are the predictor partition rule \mathcal{R}_X , the response partition rule \mathcal{R}_Y , the stopping probability $\rho(A)$, the partition selection probabilities $\lambda(A)$, and the local parameters $(\rho_Y^A, \lambda_Y^A, \alpha_Y^A)$ for all $A \subset \Omega_X$ that could arise during the predictor partition under \mathcal{R}_X .

Remark I: To ensure that this definition is meaningful, one must check that the two-stage procedure will in fact generate a well-defined conditional distribution with probability 1. To see this, first note that because the collection of all potential sets A on Ω_X that can arise during Stage I is countable, by Theorem 1 in [40], with probability 1, the two-stage procedure will generate an absolutely continuous conditional distribution of Y given $X = x$ for x in the stopped part of Ω_X , provided that ρ_Y^A is uniformly away from 0. The two-stage generation procedure for the conditional density of Y can then be completed by letting Y given X be uniform on Ω_Y for the μ_X -null subset of Ω_X on which the recursive partition in Stage I never stops.

Remark II: While the cond-OPT prior involves many hyperparameters, one can appeal to very simple symmetry and self-similarity principles to for choosing their values. Specifically, such considerations lead to the simple choice: (i) $\rho(A) \equiv \rho \in [0, 1]$, (ii) $\lambda_j(A) = 1/N(A)$, and (iii) $\rho_Y^A \equiv \rho_Y$, $\lambda_Y^A \equiv \lambda_Y$, and $\alpha_Y^A \equiv \alpha_Y$ for all A , following the default choices in [40]. We note that when useful prior knowledge about the structure of the underlying distribution is not available or when one is unwilling to assume particular structure over the distribution, it is desirable to specify the prior parameters in a symmetric and self-similar way. The common stopping probability ρ should not be too close to 0 or 1, but taking a moderate value between 0.1 and 0.9. A sensitivity analysis for such choices demonstrating the robustness of such choices in the context of OPTs is provided in [22]. As for the partition rules, the coordinate-wise dyadic mid-split rule can serve as a simple default choice for both \mathcal{R}_X and \mathcal{R}_Y . We will adopt such a specification in all of our numerical examples.

We have emphasized that the cond-OPT prior imposes no parametric assumptions on the conditional distribution. One may wonder whether this prior is truly “nonparametric” in the sense that it can generate all possible conditional densities. Our next theorem confirms this—under mild conditions on the parameters, the cond-OPT will place positive probability in arbitrarily small L_1 neighborhoods of any conditional density. (A definition of an L_1 neighborhood for conditional densities is also implied in the statement of the theorem.)

Theorem 2 (Large support). *Suppose $q(\cdot|\cdot)$ is a conditional density function that arises from a cond-OPT prior whose parameters $\rho(A)$ and $\lambda(A)$ for all A that could arise during the recursive partitioning on Ω_X are uniformly away from 0 and 1, and the local parameters satisfy the conditions specified in Theorem 2 of [40]. Moreover, suppose that the underlying partition rules \mathcal{R}_X and \mathcal{R}_Y both satisfy the following “fine partition criterion”: $\forall \epsilon > 0$, there exists a partition of the corresponding sample space such that the diameter of each partition block is less than ϵ . Then for any conditional density function $f(\cdot|\cdot) : \Omega_Y \times \Omega_X \rightarrow [0, \infty)$, and any $\tau > 0$,*

$$P\left(\int |q(y|x) - f(y|x)|\mu(dx \times dy) < \tau\right) > 0.$$

Furthermore, let $f_X(x)$ be any density function on Ω_X w.r.t. μ_X . Then we have $\forall \tau > 0$,

$$P\left(\int |q(y|x) - f(y|x)|f_X(x)\mu(dx \times dy) < \tau\right) > 0.$$

3. Bayesian inference with cond-OPT

Next we investigate how Bayesian inference on conditional densities can be carried out using this prior. First, we note that Chipman et al [2] and Denison et al [5] each proposed MCMC algorithms that enable posterior inference for Bayesian CART. These sampling and stochastic search algorithms can be applied directly here as the local OPT priors can be marginalized out and so the marginal likelihood under each partition tree that arises in Stage I of the cond-OPT is available in closed form [40, 22]. However, as noted in [2] along with others, due to the multi-modal nature of tree structured models, the mixing behavior of the MCMC algorithms is often undesirable. This problem is exacerbated in higher dimensional settings. Chipman et al [2] suggested using MCMC as a tool for searching for good models rather than a reliable way of sampling from the actual posterior.

The main result of this section is that Bayesian inference under a cond-OPT prior can be carried out in an *exact* manner, in that the corresponding posterior distribution can be computed in closed form and directly sampled from, without resorting to MCMC algorithms. Not only is the computation feasible under simple partition rules such as the coordinate-wise dyadic mid-split rule for multivariate sample spaces of moderate dimensions (e.g., ≤ 5 continuous dimensions), but it is in fact highly efficient, scaling linearly with the number of observations.

First let us investigate what the posterior of a cond-OPT prior is. Suppose we have observed $(\mathbf{x}, \mathbf{y}) = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ where given the x_i 's, the y_i 's are independent with some density $q(y|x)$. We assume that $q(\cdot|\cdot)$ has a cond-OPT prior denoted by π . Further, for any $A \subset \Omega_{\mathbf{X}}$ we let

$$\mathbf{x}(A) := \{x_1, x_2, \dots, x_n\} \cap A \quad \text{and} \quad \mathbf{y}(A) := \{y_i : x_i \in A, i = 1, 2, \dots, n\},$$

and let $n(A)$ denote the number of observations with predictors lying in A , that is $n(A) = |\mathbf{x}(A)| = |\mathbf{y}(A)|$.

For $A \subset \Omega_{\mathbf{X}}$, we use $q(A)$ to denote the (conditional) likelihood under $q(\cdot|\cdot)$ contributed from the data with predictors $x \in A$. That is

$$q(A) := \prod_{i: x_i \in A} q(y_i | x_i).$$

Then conditional on the event that A arises during the recursive partition procedure on $\Omega_{\mathbf{X}}$, we can write $q(A)$ recursively in terms of $S(A)$, $J(A)$, and $q_{\mathbf{Y}}^A$ as follows

$$q(A) = \begin{cases} q^0(A) & \text{if } S(A) = 1 \\ \prod_{i=1}^{K^J(A)} q(A_i^j) & \text{if } S(A) = 0 \text{ and } J(A) = j, \end{cases}$$

where

$$q^0(A) := \prod_{i: x_i \in A} q_{\mathbf{Y}}^{0,A}(y_i),$$

the likelihood from the data with $x \in A$ if the partitioning stops on A . Equivalently, we can write

$$q(A) = S(A)q^0(A) + (1 - S(A)) \prod_{i=1}^{K^{J(A)}(A)} q(A_i^{J(A)}). \quad (3.1)$$

Integrating out the randomness over both sides of Eq. (3.1), we get

$$\Phi(A) = \rho(A)M(A) + (1 - \rho(A)) \sum_{j=1}^{N(A)} \lambda_j(A) \prod_i \Phi(A_i^j), \quad (3.2)$$

where

$$\Phi(A) := \int q(A) \pi(dq | A \text{ arises during the recursive partitioning})$$

is defined to be the marginal likelihood from data with $x \in A$ given that A arises during the recursive partitioning on $\Omega_{\mathbf{X}}$, whereas

$$M(A) := \int q^0(A) \pi(dq_{\mathbf{Y}}^{0,A}) \quad (3.3)$$

is the marginal likelihood from the data with $x \in A$ if the recursive partitioning procedure stops on A and the integration is taken over the local $\text{OPT}(\mathcal{R}_{\mathbf{Y}}; \rho_{\mathbf{Y}}^A, \lambda_{\mathbf{Y}}^A, \alpha_{\mathbf{Y}}^A)$ prior for $q_{\mathbf{Y}}^{0,A}$. We note that Eqs. (3.1), (3.2) and (3.3) hold for Bayesian CART as well, with $M(A)$ being the corresponding marginal likelihood of the local normal model or the multinomial model under the corresponding priors such as those given earlier.

Eq. (3.2) provides a recursive recipe for calculating $\Phi(A)$ for all A . It is recursive in the sense that $\Phi(A)$ is computed based on the value of $\Phi(\cdot)$ on A 's children. (Of course, to complete the calculation the recursion must eventually terminate everywhere on $\Omega_{\mathbf{X}}$. We shall describe the terminal conditions in the next section.) This recursive algorithm is a special case of the forward-backward algorithm [21].

The next theorem establishes the posterior conjugacy of cond-OPT.

Theorem 3 (Conjugacy). *After observing $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ where given the x_i 's, the y_i 's are independent with density $q(y|x)$, which has a cond-OPT prior, the posterior of $q(\cdot|\cdot)$ is again a cond-OPT (with the same partition rules on $\Omega_{\mathbf{X}}$ and $\Omega_{\mathbf{Y}}$ as the prior). Moreover, for each $A \subset \Omega_{\mathbf{X}}$ that could arise during the recursive partitioning, the posterior parameters are given as follows.*

1. *Stopping probability:*

$$\rho(A|\mathbf{x}, \mathbf{y}) = \rho(A)M(A)/\Phi(A).$$

2. *Selection probabilities:*

$$\lambda_j(A|\mathbf{x}, \mathbf{y}) = \lambda_j(A) \frac{(1 - \rho(A)) \prod_{i=1}^{K^j(A)} \Phi(A_i^j)}{\Phi(A) - \rho(A)M(A)}.$$

3. *The local parameters: $\tilde{\rho}_{\mathbf{Y}}^A$, $\tilde{\lambda}_{\mathbf{Y}}^A$, and $\tilde{\alpha}_{\mathbf{Y}}^A$ are the corresponding posterior parameters for the local OPT after updating using the observed values for the response $\mathbf{y}(A)$, $\text{OPT}(\mathcal{R}_{\mathbf{Y}}^A; \tilde{\rho}_{\mathbf{Y}}^A, \tilde{\lambda}_{\mathbf{Y}}^A, \tilde{\alpha}_{\mathbf{Y}}^A)$.*

This theorem shows that *a posteriori* our knowledge about the underlying conditional distribution of \mathbf{Y} given \mathbf{X} can again be represented by the same two-stage procedure that randomly partitions the predictor space and then generate the response distribution accordingly on each of the predictor block, except that now the parameters that characterize this two-stage procedure have been updated to reflect the information contained in the data. Moreover, the theorem also provides a recipe for computing these posterior parameters based on $\Phi(A)$ and $M(A)$. Given this exact posterior, Bayesian inference can then proceed—samples can be drawn from the posterior cond-OPT directly through vanilla Monte Carlo (as opposed to MCMC) and summary statistics calculated.

In the next section, we provide more details on how to implement such inference in practice. Before that, we present our last theoretical result about the cond-OPT prior—its posterior consistency, which assures the statistician that the posterior cond-OPT distribution will “converge” in some sense to the truth as the amount of data increases. To this end, we first need a notion of neighborhoods for conditional densities under which such convergence holds. We adopt the notion discussed in [29] and [28], by which a (weak) neighborhood of a conditional density function is defined in terms of a (weak) neighborhood of the corresponding joint density. More specifically, for a conditional density function $f_0(\cdot|\cdot) : \Omega_{\mathbf{Y}} \times \Omega_{\mathbf{X}} \rightarrow [0, \infty)$, weak neighborhoods with respect to a marginal density $f_{\mathbf{X}}^0(\cdot)$ on $\Omega_{\mathbf{X}}$ are collections of conditional densities of the form

$$U = \left\{ f(\cdot|\cdot) : \left| \int g_i f(\cdot|\cdot) f_{\mathbf{X}}^0 d\mu - \int g_i f_0(\cdot|\cdot) f_{\mathbf{X}}^0 d\mu \right| < \epsilon_i, i = 1, 2, \dots, l \right\}$$

where the g_i 's are bounded continuous functions on $\Omega_{\mathbf{X}} \times \Omega_{\mathbf{Y}}$.

Theorem 4 (Weak consistency). *Let $(x_1, y_1), (x_2, y_2), \dots$ be independent identically distributed vectors from a probability distribution on $\Omega_{\mathbf{X}} \times \Omega_{\mathbf{Y}}$, F , with density $dF/d\mu = f(x, y) = f(y|x)f_{\mathbf{X}}(x)$. Suppose the conditional density $f(\cdot|\cdot)$ is generated from a cond-OPT prior for which the conditions given in Theorem 2*

all hold. In addition, assume that the conditional density function $f(\cdot|\cdot)$ and the joint density $f(\cdot, \cdot)$ are bounded. Then for any weak neighborhood of $f(\cdot|\cdot)$ w.r.t $f_{\mathbf{X}}, U$, we have

$$\pi(U|(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)) \longrightarrow 1$$

with F^∞ probability 1, where $\pi(\cdot|(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n))$ denotes the cond-OPT posterior for $f(\cdot|\cdot)$.

4. Practical implementation

Next we address some practical issues in computing the posterior and implementing the inference. For simplicity, from now on we shall refer to a set $A \subset \Omega_{\mathbf{X}}$ that can arise during the (Stage I) recursive partitioning procedure as a “node” (i.e., as a node in the partition tree).

A prerequisite for applying Theorem 3 is the availability of the $\Phi(A)$ terms, which can be determined recursively through Eq. (3.2). Of course, to carry out the computation of $\Phi(A)$ one must specify terminal conditions on Eq. (3.2), or in other words, on what kind of A ’s the recursion should terminate. We call such nodes *terminal* nodes.

There are two kinds of nodes for which the value of $\Phi(A)$ is available directly according to theory, and thus recursion can terminate on them. They are (i) nodes that cannot be further divided under the partition rule $\mathcal{R}_{\mathbf{X}}$, and (ii) nodes that contain no more than one data point. For a node A that cannot be further divided, we must have $\rho(A) = 1$ and so $\Phi(A) = M(A)$. For a node A with no data point, it has no contribution to the likelihood and so $\Phi(A) = 1$. For a node A with exactly one data point, $\Phi(A)$ is the predictive density of the local OPT on A evaluated at that data point, which is exactly the density of the prior mean of the OPT, or the so-called base measure, which is directly known when the default symmetric and self-similar prior specification for the local OPTs is adopted as recommended in [40].

Note that with these two types of “theoretical” terminal nodes, in principle the recursion will eventually terminate if one divides the predictor space deep enough. In practice, however, it is unnecessary to take the recursion all the way down to these theoretical terminal nodes. Instead, one can adopt early termination by imposing a technical limit, such as a minimum size of the nodes either in terms of the nature measure $\mu_{\mathbf{X}}(A)$ or the number of observations therein $n(A)$, to end the recursion. Nodes that are smaller than the chosen size threshold are forced to be terminal, which is equivalent to setting $\rho(A) = 1$ and thus $\Phi(A) = M(A)$ for these nodes. We call these nodes “technical” terminal nodes.

With these theoretical and technical terminal nodes, one can then compute $\Phi(A)$ through the recursion formula (3.2), and compute the posterior according to Theorem 3. Putting all the pieces together, we can summarize the procedure to carry out Bayesian inference with the cond-OPT prior as a four-step recipe:

- I. For each terminal node A , compute $M(A)$ and $\Phi(A)$.
- II. For each non-terminal node A (those that are ancestors of the terminal nodes), compute $M(A)$ and use Eq. (3.2) to recursively compute $\Phi(A)$.
- III. Given the values of $M(A)$ and $\Phi(A)$, apply Theorem 3 to get the parameter values of the posterior cond-OPT distribution.
- IV. Sample from the exact posterior by direct simulation of the random two-stage procedure, and/or compute summary statistics of the posterior.

For the last step, direct simulation from the posterior is straight-forward, but we have not discussed what summary statistics to compute and how to do that. This is problem-specific and will be illustrated in our numeric examples in Section 5.

5. Examples

In this section we provide four examples to illustrate inference using the cond-OPT prior. The first two illustrate the estimation of conditional densities, the third is on model selection, and the last is for hypothesis testing (in particular testing independence). In these examples, the partition rules used on both $\Omega_{\mathbf{X}}$ and $\Omega_{\mathbf{Y}}$ are always the coordinate-wise dyadic mid-split rule. We adopt the same prior specification across all the

examples: the prior stopping probability on each non-terminal node is always set to 0.5, the prior partition selection probability is always evenly spread over the possible ways to partition each set, and the probability assignment pseudo-counts for the local OPTs are all set to 0.5, and nodes at 14 levels down the partition tree, i.e., with $\mu_{\mathbf{X}}(A) = \mu(\Omega_{\mathbf{X}})/2^{14}$, are set to be the technical terminal nodes.

Example 1 (Estimating conditional density with sharp predictor boundaries). In this example we simulate (X, Y) pairs according to the following distributions.

$$\begin{aligned} X &\sim \text{Beta}(2, 2) \\ Y|X < 0.25 &\sim \text{Beta}(30, 20) \\ Y|0.25 \leq X \leq 0.5 &\sim \text{Beta}(10, 30) \\ Y|X > 0.5 &\sim \text{Beta}(0.5, 0.5). \end{aligned}$$

We generate data sets of three different sample sizes, $n = 100$, $n = 500$, and $n = 2,500$, and place the cond-OPT prior on the distribution of Y given X . Following the four-step recipe given in the previous section, we can compute the posterior cond-OPT and sample from it.

A representative summary of the posterior partitioning mechanism is the so-called hierarchical *maximum a posteriori* (hMAP) [40] partition tree, which can be computed from the posterior analytically [40] and is plotted in Figure 1 for the different sample sizes. (Chipman et al [2] and Wong and Ma [40] both discussed reasons why the commonly adopted MAP is not a good summary for tree-structured posteriors due to their multi-level nature. See [40, Sec. 4.2] for further details and reasons why the hMAP is preferred.)

In Figure 1, within each “leaf” node we plot the corresponding posterior mean of the conditional density of Y . Also plotted for each node is the posterior stopping probability. Even with only 100 data points, the posterior suggests that $\Omega_{\mathbf{X}}$ should be divided into three pieces— $[0, 0.25]$, $[0.25, 0.5]$, and $[0.5, 1]$ —within which the conditional distribution of $Y|X$ is homogeneous across X . Note that the posterior stopping probabilities on those three intervals are large, in contrast to the near 0 values on the larger sets. Reliably estimating the actual conditional density function on these sets nonparametrically appears to require more than 100 data points. In this example, a sample size of 500 already does a decent job.

The previous example favors our method because (1) there are a small number of clear boundaries of change for the underlying conditional distribution—namely 0.25 and 0.5, and (2) those boundaries lie on the potential partition points of the partition rule. In the next example, we examine the case in which the conditional distribution changes smoothly across a continuous X without any boundary of abrupt change.

Example 2 (Estimating conditional density without predictor boundaries). In this example we generate (X, Y) from a bivariate normal distribution.

$$(X, Y)' \sim \text{BN}\left(\begin{pmatrix} 0.6 \\ 0.4 \end{pmatrix}, \begin{pmatrix} 0.1^2 & 0.005 \\ 0.005 & 0.1^2 \end{pmatrix}\right).$$

We generate data sets of size $n = 2,000$, and apply the cond-OPT prior on the distribution of Y given X as we did in the previous example. Again we compute the posterior cond-OPT following our four-step recipe. The hMAP tree and the posterior mean estimate of the conditional densities are presented in Figure 2. Because the underlying predictor space Ω_X is unbounded, for simplicity we use the empirically observed range of X as Ω_X , which happens to be $\Omega_X = [0.24, 0.92]$ for our simulated example. (Other ways to handle this situation include transforming X to have a compact support such as through a CDF or rank transform.

As illustrated in the figure, as the sample size increases, the posterior cond-OPT will partition Ω_X into finer blocks, reflecting the fact that the conditional density changes smoothly across $\Omega_{\mathbf{X}}$. One interesting observation is that the “leaf” nodes in Figure 2 have very large (close to 1) posterior stopping probability. This may seem surprising as the underlying conditional distribution is not the same for any neighboring values of X . The large posterior stopping probabilities indicate that on those sets, where the sample size is not large, the gain in achieving better estimate of the common features of the conditional distribution for nearby X values outweighs the loss in ignoring the difference among them.

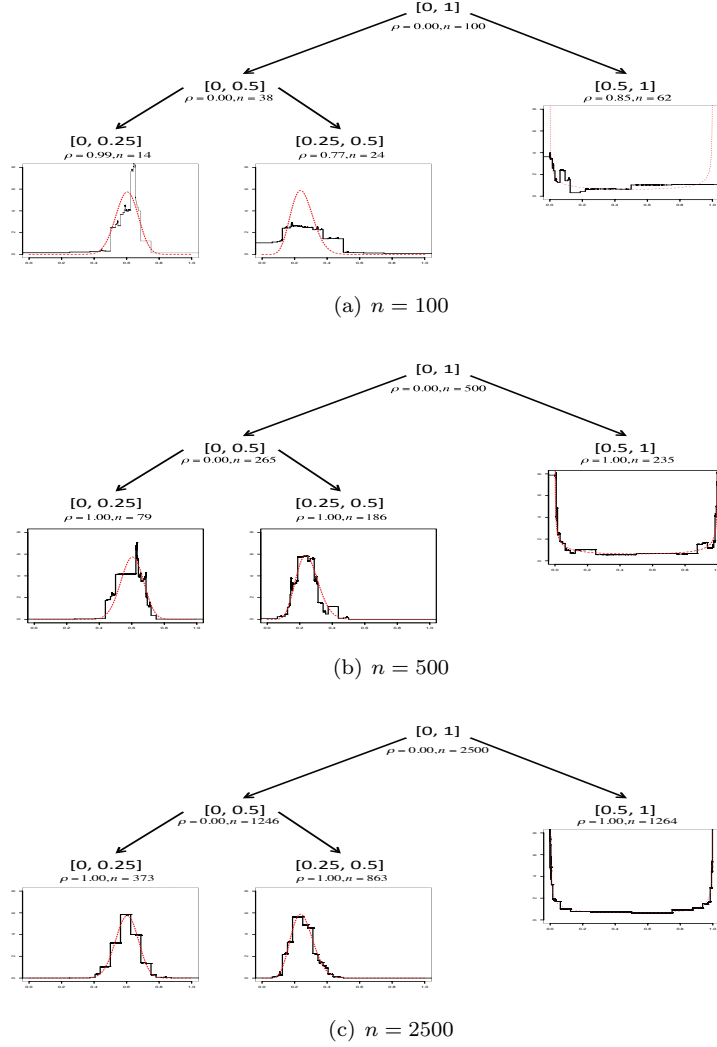


FIG 1. The hMAP partition tree structures on X and the posterior mean estimate of $Y|X$ given the tree structure for Example 1. For each node, ρ indicates the posterior stopping probability for each node and n represents the number of data points in each node. The plot under each stopped node gives the mean of the posterior local OPT for Y within that node (solid line) along with the true conditional densities (dashed line).

Example 3 (Model selection over binary predictors). Next we show how one can use cond-OPT to carry out model selection. Consider the case in which $\mathbf{X} = (X_1, X_2, \dots, X_{30}) \in \{0, 1\}^{30}$ forming a Markov Chain:

$$X_1 \sim \text{Bernoulli}(0.5) \quad \text{and} \quad P(X_i = X_{i-1} | X_{i-1}) = 0.7$$

for $i = 2, 3, \dots, 30$. Suppose the conditional distribution of a continuous response Y is

$$Y \sim \begin{cases} \text{Beta}(1, 6) & \text{if } (X_5, X_{20}, X_{30}) = (1, 0, 1) \\ \text{Beta}(12, 16) & \text{if } (X_5, X_{20}) = (0, 1) \\ \text{Beta}(3, 4) & \text{otherwise.} \end{cases}$$

In other words, three predictors X_5 , X_{20} and X_{30} impact the response in an interactive manner. Our interest is in recovering this underlying interactive structure (i.e. the “model”). To illustrate, we simulate 500 data points from this scenario and place a cond-OPT prior on $Y|\mathbf{X}$, and consider predictor partitions up to four

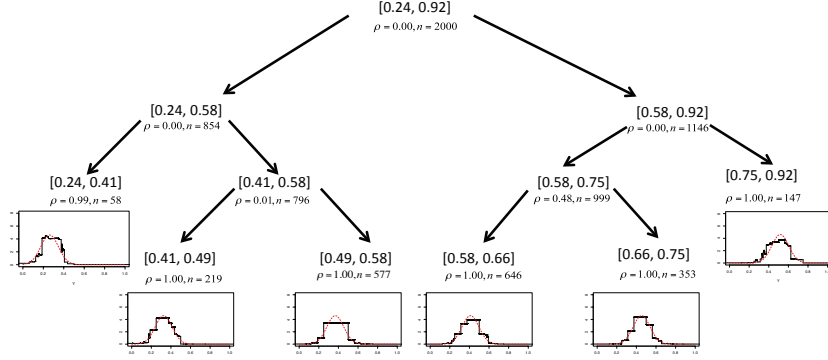


FIG 2. The hMAP trees on Ω_X and the posterior mean estimate of $Y|X$ within the stopped sets for Example 2. The plot under each stopped node gives the mean of the posterior local OPT for Y within that node (solid line) along with the true conditional densities at the center value of the stopped predictor intervals (dashed line). For each node, ρ indicates the posterior stopping probability for each node and n represents the number of data points in each node.

levels deep. This is achieved by setting $\rho(A) = 1$ for A that arises after four steps of partitioning, and it allows us to search for models involving up to four-way interactions. We again carry out the four-step recipe to get the posterior and calculate the hMAP. The hMAP tree structure along with the predictive density for $Y|X$ within each stopped set is presented in Figure 3. The posterior concentrates on partitions involving X_5 , X_{20} and X_{30} out of the 30 variables. While the predictive conditional density for $Y|X$ is very rough given the limited number of data points in the stopped sets, the posterior recovers the exact interactive structure of the predictors with little uncertainty.

In addition, we sample from the posterior and use the proportion of times each predictor appears in the sampled models to estimate the posterior marginal inclusion probabilities. Our estimates based on 1,000 draws from the posterior are presented in Figure 4(a). Note that the sample size 500 is so large that the posterior marginal inclusion probabilities for the three relevant predictors are all close to 1 while those for the other predictors are close to 0. We carry out the same simulation with a reduced sample size of 200, and plot the estimated posterior marginal inclusion probabilities in Figure 4(b). We see that with a sample size of 200, one can already use the posterior to reliably recover the relevant predictors.

Example 4 (Test of independence). In this example, we illustrate an application of the cond-OPT prior for hypothesis testing. In particular, we use it to test the independence between \mathbf{X} and \mathbf{Y} . To begin, note that $\rho(A|\mathbf{x}, \mathbf{y})$ in Theorem 3 gives the posterior probability for the conditional distribution of \mathbf{Y} to be constant over all values of \mathbf{X} in A , or in other words, for \mathbf{Y} to be independent of \mathbf{X} on A . Hence, one can consider $\rho(\Omega_X|\mathbf{x}, \mathbf{y})$ as a score for the statistical significance of dependence between the observed variables. A permutation null distribution of this statistic can be constructed by randomly pairing the observed \mathbf{x} and \mathbf{y} values, and based on this, p-values can be computed for testing the null hypothesis of independence.

To illustrate, we simulate $\mathbf{X} = (X_1, X_2, \dots, X_{10})$ for a sample of size 400 under the same Markov Chain model as in the previous example, and simulate a response variable Y as follows.

$$Y \sim \begin{cases} \text{Beta}(4, 4) & \text{if } (X_1, X_2, X_5) = (1, 1, 0) \\ \text{Beta}(0.5, 0.5) & \text{if } (X_5, X_8, X_{10}) = (1, 0, 0) \\ \text{Unif}[0, 1] & \text{otherwise.} \end{cases}$$

In particular, \mathbf{Y} is dependent on \mathbf{X} but there is no mean or median shift in the conditional distribution of Y over different values of \mathbf{X} . Figure 5 gives the histogram of $\rho(\Omega_X|\mathbf{x}, \mathbf{y})$ for 1,000 permuted samples where the vertical dashed line indicates the $\rho(\Omega_X|\mathbf{x}, \mathbf{y})$ for the original simulated data, which equals 0.0384. For this particular simulation, 7 out of the 1,000 permuted samples produced a more extreme test statistic.

Remark I: Note that by symmetry one can place a cond-OPT prior on the conditional distribution of \mathbf{X}

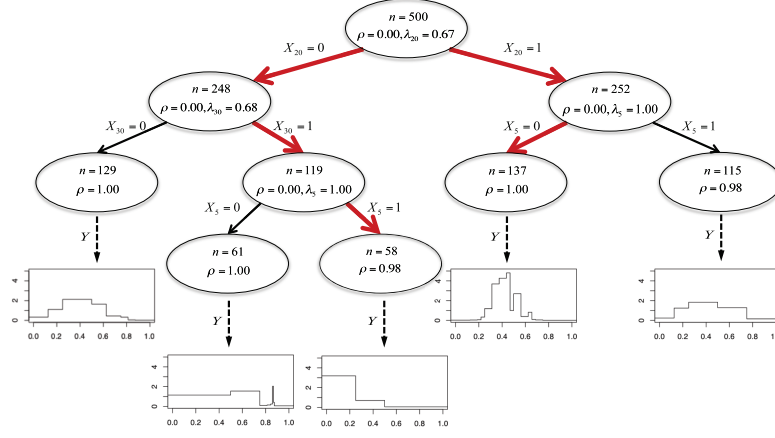


FIG 3. The hMAP tree structure on $\Omega_{\mathbf{X}}$ and the posterior mean estimate of $Y|\mathbf{X}$ in each of the stopped sets for Example 3. The bold arrows indicate the “true model”—predictor combinations that correspond to “non-null” $Y|\mathbf{X}$ distributions. For each node, ρ indicates the posterior stopping probability for each node, λ represents the posterior selection probability for the most probable direction if the partition does not stop on the node, and n represents the number of data points in each node.

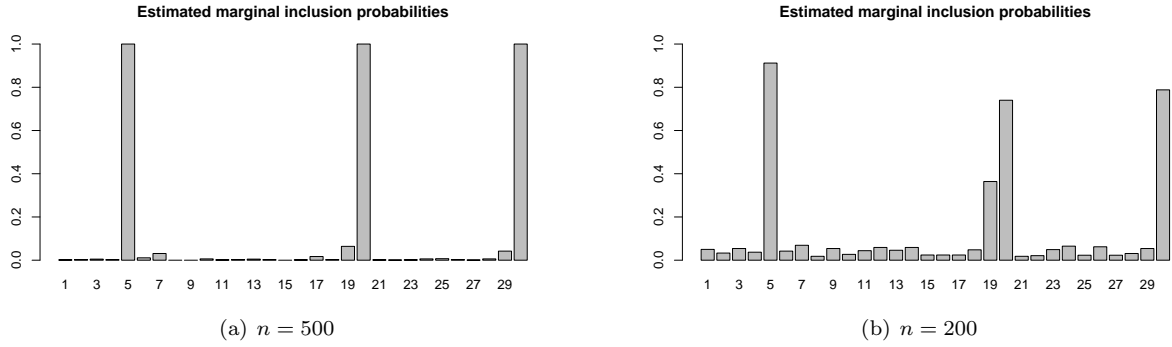


FIG 4. Estimated posterior marginal inclusion probabilities for the 30 predictors in Example 3 for two different sample sizes. The estimates are computed over 1,000 draws from the corresponding posteriors.

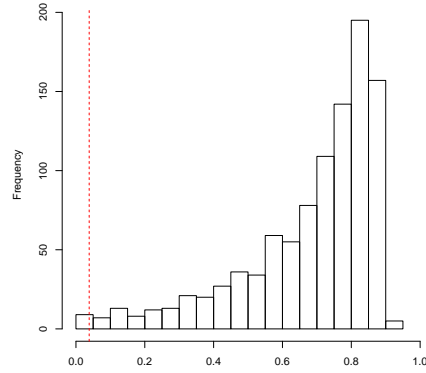


FIG 5. Histogram of $\rho(\Omega_{\mathbf{X}}|\mathbf{x}, \mathbf{y})$ for 1,000 permuted samples. The vertical line indicates $\rho(\Omega_{\mathbf{X}}|\mathbf{x}, \mathbf{y})$ for the original data.

given \mathbf{Y} as well and that will produce a corresponding posterior stopping probability $\rho(\Omega_{\mathbf{Y}}|\mathbf{y}, \mathbf{x})$. One can thus alternatively use $\min\{\rho(\Omega_{\mathbf{X}}|\mathbf{x}, \mathbf{y}), \rho(\Omega_{\mathbf{Y}}|\mathbf{y}, \mathbf{x})\}$ as the test statistic for independence.

Remark II: Testing using the posterior stopping probability $\rho(\Omega_{\mathbf{X}}|\mathbf{x}, \mathbf{y})$ is equivalent to using a Bayes factor (BF). To see this, note that the BF for testing independence under the cond-OPT can be written as

$$\text{BF}_{\mathbf{Y}|\mathbf{X}} = \frac{\sum_{j=1}^{N(A)} \lambda_j(A) \prod_i \Phi(A_i^j)}{M(A)}$$

with $A = \Omega_{\mathbf{X}}$ where the numerator is the marginal conditional likelihood of \mathbf{Y} given \mathbf{X} if the conditional distribution of \mathbf{Y} is not constant over \mathbf{X} (i.e. $\Omega_{\mathbf{X}}$ is divided) and the denominator is that if the conditional distribution of \mathbf{Y} is the same for all \mathbf{X} (i.e. $\Omega_{\mathbf{X}}$ is undivided). By Eq. (3.2) and Theorem 3,

$$\text{BF}_{\mathbf{Y}|\mathbf{X}} = \frac{\rho(\Omega_{\mathbf{X}})}{1 - \rho(\Omega_{\mathbf{X}})} \left(\frac{1}{\rho(\Omega_{\mathbf{X}}|\mathbf{x}, \mathbf{y})} - 1 \right),$$

which is in a one-to-one correspondence to $\rho(\Omega_{\mathbf{X}}|\mathbf{x}, \mathbf{y})$ given the prior parameters.

6. Application to real data: conditional density estimation in flow cytometry

In flow cytometry experiments for immunological studies, a number (typically 4 to 10) of biomarkers are measured on large numbers of blood cells. Estimated densities and conditional densities of such data can be used for tasks such as automatic classification of the cells [24]. We apply cond-OPT to estimate the conditional density of one marker “CD4” given another “CD8” in a flow cytometry data set. This particular data set contains $n = 455,472$ cells are measured. Flow cytometry experiments often involve large numbers of cells, and thus practical methods must scale well in computing time and memory usage with respect to the number of observations. This poses great challenge to existing nonparametric models that require intense MCMC computation. The values of the two markers are measured in the range of $[0,1]$. We use exactly the same technical specifications of the cond-OPT prior as in our simulation examples.

Figure 6 presents the posterior mean of the conditional density of CD4 given CD8 under the cond-OPT model given the hMAP partition on the predictor space, which splits the space into 34 pieces. The entire computation of the full posterior, the hMAP partition, as well as the conditional posterior expectation of the conditional density given the hMAP tree, took about 15 seconds to complete on a single 3.6GHz Intel Core-i7 3820 desktop core without parallelization and required about 5 Gbs of memory.

7. Discussion

In this work we have introduced a Bayesian nonparametric prior on the space of conditional densities. This prior, which we call the conditional optional Pólya tree, is constructed based on a two-stage procedure that first divides the predictor space $\Omega_{\mathbf{X}}$ and then generates the conditional distribution of the response through local OPT processes. We have established several important theoretical properties of this prior, namely large support, conjugacy and posterior consistency, and have provided a practical recipe for Bayesian inference using this prior.

The construction of this prior does not depend on the marginal distribution of \mathbf{X} . One particular implication is that one can transform \mathbf{X} before applying the prior on $\mathbf{Y}|\mathbf{X}$ without invalidating the posterior inference. (Note that transforming \mathbf{X} is equivalent to choosing a different partition rule on $\Omega_{\mathbf{X}}$.) In certain situations it is desirable to perform such a transformation on \mathbf{X} . For example, if the data points are very unevenly spread over $\Omega_{\mathbf{X}}$, then some parts of the space may contain a very small number of data points. There the posterior is mostly dominated by the prior specification and does not provide much information about the underlying conditional distribution. One way to mitigate this problem is to transform \mathbf{X} so that the data are more evenly distributed over $\Omega_{\mathbf{X}}$. When $\Omega_{\mathbf{X}}$ is one-dimensional, for example, this can be achieved by a rank transformation on X . Another situation in which a transformation of \mathbf{X} may be useful is when the dimensionality of \mathbf{X} is very high. In this case a dimensionality reduction transformation can be applied

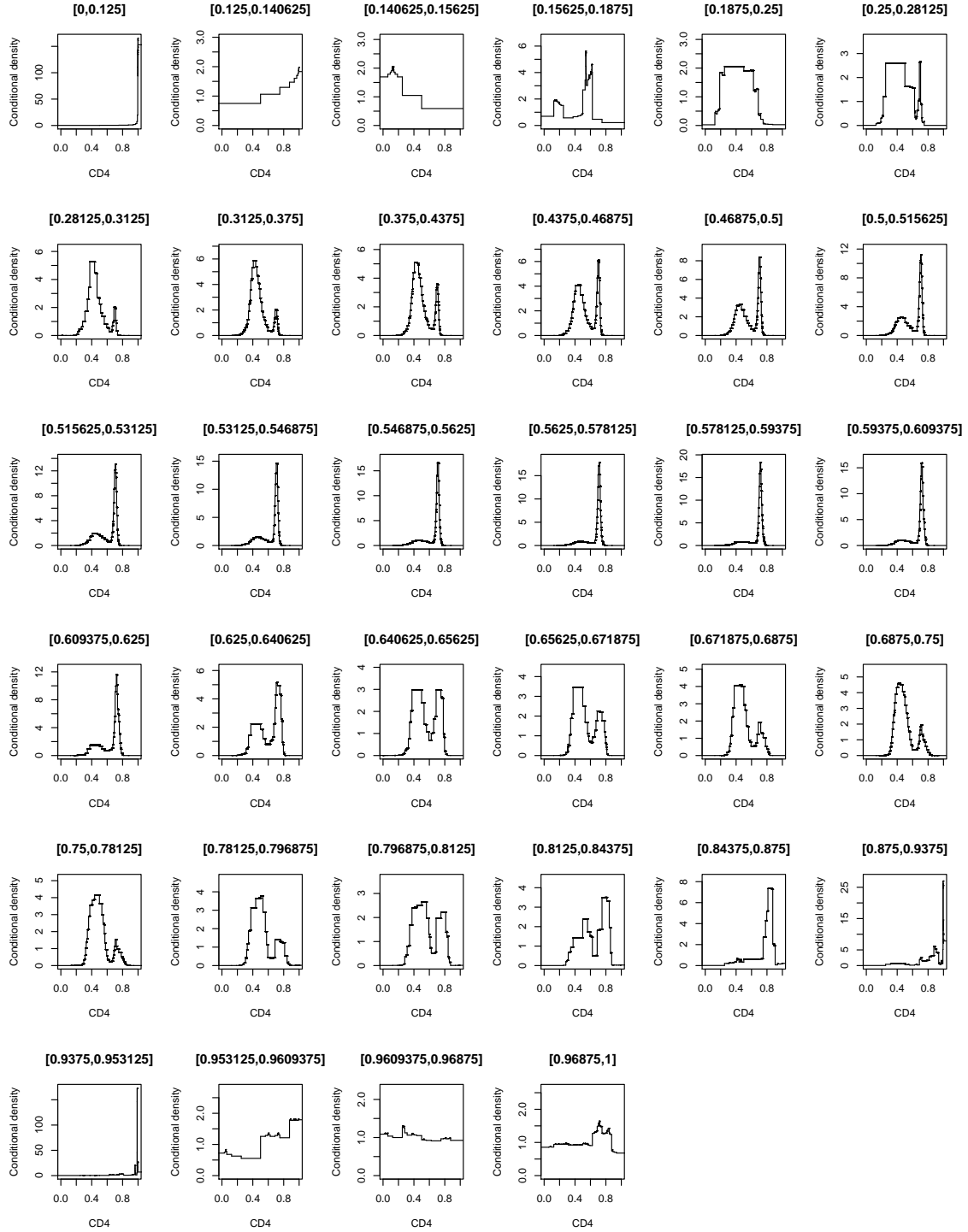


FIG 6. The posterior mean conditional densities of the CD4 marker given the CD8 marker conditional on the hMAP partition on the CD8 values for the flow cytometry data set. The range on top of each plot gives the corresponding block in the hMAP partition of the predictor (CD8) space.

on \mathbf{X} before carrying out the inference. Of course, in doing so one often loses the ability to interpret the posterior conditional distribution of \mathbf{Y} directly in terms of the original predictors.

Finally, while we have used recursive partitioning in conjunction with the OPT to build a model for conditional density, one can build such models by replacing the OPT with other multi-scale density models in the PT family, such as the more recently introduced adaptive Pólya tree (APT) [21].

Acknowledgment

The flow cytometry data set was provided by EQAPOL (HHSN272201000045C), an NIH/NIAID/DAIDS-sponsored, international resource that supports the development, implementation, and oversight of quality assurance programs (Sanchez PMC4138253).

Appendix: Proofs

Proof of Lemma 1. The proof of this lemma is very similar to that of Theorem 1 in [40]. Let T_1^k be the part of $\Omega_{\mathbf{X}}$ that has not been stopped after k levels of recursive partitioning. The random partition of $\Omega_{\mathbf{X}}$ after k levels of recursive partitioning can be thought of as being generated in two steps. First suppose there is no stopping on any set and let $J^{*(k)}$ be the collection of partition selection variables J generated in the first k levels of recursive partitioning. Let $\mathcal{A}^k(J^{*(k)})$ be the collection of sets A that arise in the first k levels of non-stopping recursive partitioning, which is determined by $J^{*(k)}$. Then we generate the stopping variables $S(A)$ for each $A \in \mathcal{A}^k(J^{*(k)})$ successively for $k = 1, 2, \dots$, and once a set is stopped, let all its descendants be stopped as well. Now for each $A \in \mathcal{A}^k(J^{*(k)})$, let $I^k(A)$ be the indicator for A 's stopping status after k levels of recursive partitioning, with $I^k(A) = 1$ if A is not stopped and $= 0$ otherwise.

$$\begin{aligned} E(\mu_{\mathbf{X}}(T_1^k)|J^{*(k)}) &= E\left(\sum_{A \in \mathcal{A}^k(J^{*(k)})} \mu_{\mathbf{X}}(A) I^k(A) | J^{*(k)}\right) \\ &= \sum_{A \in \mathcal{A}^k(J^{*(k)})} \mu_{\mathbf{X}}(A) E(I^k(A) | J^{*(k)}) \\ &\leq \mu_{\mathbf{X}}(\Omega_{\mathbf{X}})(1 - \delta)^k. \end{aligned}$$

Hence $E(\mu_{\mathbf{X}}(T_1^k)) \leq \mu_{\mathbf{X}}(\Omega_{\mathbf{X}})(1 - \delta)^k$, by Markov inequality and Borel-Contelli lemma, we have $\mu_{\mathbf{X}}(T_1^k) \downarrow 0$ with probability 1. \square

Proof of Theorem 2. We prove only the second result as the first follows by choosing $f_{\mathbf{X}}(x) \equiv 1/\mu_{\mathbf{X}}(\Omega_{\mathbf{X}})$. Also, we consider only the case when $\Omega_{\mathbf{X}}$ and $\Omega_{\mathbf{Y}}$ are both compact Euclidean rectangles, because the cases when at least one of the two spaces is finite follow as simpler special cases. For $x \in \Omega_{\mathbf{X}}$ and $y \in \Omega_{\mathbf{Y}}$, let $f(x, y) := f_{\mathbf{X}}(x)f(y|x)$ denote the joint density. First we assume that the joint density $f(x, y)$ is uniformly continuous. In this case it is bounded on $\Omega_{\mathbf{X}} \times \Omega_{\mathbf{Y}}$. We let $M := \sup f(x, y)$ and

$$\delta(\epsilon) := \sup_{|x_1 - x_2| + |y_1 - y_2| < \epsilon} |f(x_1, y_1) - f(x_2, y_2)|.$$

By uniform continuity, we have $\delta(\epsilon) \downarrow 0$ as $\epsilon \downarrow 0$. In addition, we define

$$\begin{aligned} \delta_{\mathbf{X}}(\epsilon) &:= \sup_{|x_1 - x_2| < \epsilon} |f_{\mathbf{X}}(x_1) - f_{\mathbf{X}}(x_2)| \\ &\leq \int \sup_{|x_1 - x_2| < \epsilon} |f(x_1, y) - f(x_2, y)| \mu_{\mathbf{Y}}(dy) \leq \delta(\epsilon) \mu_{\mathbf{Y}}(\Omega_{\mathbf{Y}}). \end{aligned}$$

Note that in particular the continuity of $f(x, y)$ implies the continuity of $f_{\mathbf{X}}(x)$. Let $\sigma > 0$ be any positive constant. Choose a positive constant $\epsilon(\sigma)$ such that $\delta_{\mathbf{X}}(\epsilon(\sigma)) = \delta(\epsilon(\sigma)) \mu_{\mathbf{Y}}(\Omega_{\mathbf{Y}}) < \max(\sigma/2, \sigma^3/2)$. Because all the parameters in the cond-OPT are uniformly bounded away from 0 and 1, there is positive probability

that $\Omega_{\mathbf{X}}$ will be partitioned into $\Omega_{\mathbf{X}} = \cup_{i=1}^K B_i$ where the diameter of each B_i is less than $\epsilon(\sigma)$, and the partition stops on each of the B_i 's. (The existence of such a partition follows from the fine partition criterion.) Let $A_i = B_i \cap \{\mathbf{X} : f_{\mathbf{X}}(x) \geq \sigma\}$, $P(\mathbf{X} \in A_i) = \int_{A_i} f_{\mathbf{X}}(x) \mu_{\mathbf{X}}(dx)$, and $f_i(y) := \int_{A_i} f(x, y) \mu_{\mathbf{X}}(dx) / \mu_{\mathbf{X}}(A_i)$ if $\mu_{\mathbf{X}}(A_i) > 0$, and 0 otherwise. Let $\mathcal{I} \subset \{1, 2, \dots, K\}$ be the set of indices i such that $\mu_{\mathbf{X}}(A_i) > 0$. Then

$$\begin{aligned} & \int |q(y|x) - f(y|x)| f_{\mathbf{X}}(x) \mu(dx \times dy) \\ & \leq \int_{f_{\mathbf{X}}(x) < \sigma} |q(y|x) - f(y|x)| f_{\mathbf{X}}(x) \mu(dx \times dy) \\ & \quad + \sum_{i \in \mathcal{I}} \int_{A_i \times \Omega_{\mathbf{Y}}} \left| q(y|x) - f_i(y) \cdot \frac{\mu_{\mathbf{X}}(A_i)}{P(\mathbf{X} \in A_i)} \right| f_{\mathbf{X}}(x) \mu(dx \times dy) \\ & \quad + \sum_{i \in \mathcal{I}} \int_{A_i \times \Omega_{\mathbf{Y}}} f_i(y) \left| \frac{\mu_{\mathbf{X}}(A_i)}{P(\mathbf{X} \in A_i)} - \frac{1}{f_{\mathbf{X}}(x)} \right| f_{\mathbf{X}}(x) \mu(dx \times dy) \\ & \quad + \sum_{i \in \mathcal{I}} \int_{A_i \times \Omega_{\mathbf{Y}}} |f_i(y) - f(x, y)| \mu(dx \times dy). \end{aligned}$$

Let us consider each of the four terms on the right hand side in turn. First,

$$\int_{f_{\mathbf{X}}(x) < \sigma} |q(y|x) - f(y|x)| f_{\mathbf{X}}(x) \mu(dx \times dy) \leq 2\sigma \mu_{\mathbf{X}}(\Omega_{\mathbf{X}}).$$

Note that for each $i \in \mathcal{I}$, $f_i(y) \mu_{\mathbf{X}}(A_i) / P(\mathbf{X} \in A_i)$ is a density function in y . Therefore by the large support property of the OPT prior (Theorem 2 in [40]), with positive probability,

$$\int_{\Omega_{\mathbf{Y}}} \left| q_{\mathbf{Y}}^{0, B_i}(y) - f_i(y) \cdot \frac{\mu_{\mathbf{X}}(A_i)}{P(\mathbf{X} \in A_i)} \right| \mu_{\mathbf{Y}}(dy) < \sigma,$$

and so

$$\int_{A_i \times \Omega_{\mathbf{Y}}} \left| q(y|x) - f_i(y) \cdot \frac{\mu_{\mathbf{X}}(A_i)}{P(\mathbf{X} \in A_i)} \right| f_{\mathbf{X}}(x) \mu(dx \times dy) < \sigma P(\mathbf{X} \in A_i)$$

for all $i \in \mathcal{I}$. Also, for any $x \in A_i$, by the choice of $\epsilon(\sigma)$,

$$\left| \frac{\mu_{\mathbf{X}}(A_i)}{P(\mathbf{X} \in A_i)} - \frac{1}{f_{\mathbf{X}}(x)} \right| \leq \frac{\delta_{\mathbf{X}}(\epsilon(\sigma))}{\sigma(\sigma - \delta_{\mathbf{X}}(\epsilon(\sigma)))} \leq \frac{\sigma^3/2}{\sigma^2/2} = \sigma.$$

Thus

$$\int_{A_i \times \Omega_{\mathbf{Y}}} f_i(y) \left| \frac{\mu_{\mathbf{X}}(A_i)}{P(\mathbf{X} \in A_i)} - \frac{1}{f_{\mathbf{X}}(x)} \right| f_{\mathbf{X}}(x) \mu(dx \times dy) \leq \sigma M \mu_{\mathbf{Y}}(\Omega_{\mathbf{Y}}) P(\mathbf{X} \in A_i).$$

Finally, again by the choice of $\epsilon(\sigma)$, $|f_i(y) - f(x, y)| \leq \delta(\epsilon(\sigma)) < \sigma$, and so

$$\int_{A_i \times \Omega_{\mathbf{Y}}} |f_i(y) - f(x, y)| \mu(dx \times dy) < \sigma \mu_{\mathbf{Y}}(\Omega_{\mathbf{Y}}) \mu_{\mathbf{X}}(A_i).$$

Therefore for any $\tau > 0$, by choosing a small enough σ , we can have

$$\int |q(y|x) - f(y|x)| f_{\mathbf{X}}(x) \mu(dx \times dy) < \tau$$

with positive probability. This completes the proof of the theorem for continuous $f(x, y)$. Now we can approximate any density function $f(x, y)$ arbitrarily close in L_1 distance by a continuous one $\tilde{f}(x, y)$. The

theorem still holds because

$$\begin{aligned}
\int |q(y|x) - f(y|x)| f_{\mathbf{X}}(x) \mu(dx \times dy) &\leq \int q(y|x) |f_{\mathbf{X}}(x) - \tilde{f}_{\mathbf{X}}(x)| \mu(dx \times dy) \\
&\quad + \int |q(y|x) - \tilde{f}(y|x)| \tilde{f}_{\mathbf{X}}(x) \mu(dx \times dy) \\
&\quad + \int |\tilde{f}(x, y) - f(x, y)| \mu(dx \times dy). \\
&\leq \int |q(y|x) - \tilde{f}(y|x)| \tilde{f}_{\mathbf{X}}(x) \mu(dx \times dy) \\
&\quad + 2 \int |\tilde{f}(x, y) - f(x, y)| \mu(dx \times dy),
\end{aligned}$$

where $\tilde{f}_{\mathbf{X}}(x)$ and $\tilde{f}(y|x)$ denote the corresponding marginal and conditional density functions for $\tilde{f}(x, y)$. \square

Proof of Theorem 3. Given that a set A is reached during the random partitioning steps on $\Omega_{\mathbf{X}}$, $\Phi(A)$ is the marginal conditional likelihood of

$$\{\mathbf{Y}(A) = \mathbf{y}(A)\} \text{ given } \{\mathbf{X}(A) = \mathbf{x}(A)\}.$$

The first term on the right hand side of Eq. (3.2), $\rho(A)M(A)$, is the marginal conditional likelihood of

$$\{\text{Stop partitioning on } A, \mathbf{Y}(A) = \mathbf{y}(A)\} \text{ given } \{\mathbf{X}(A) = \mathbf{x}(A)\}.$$

Each summand in the second term, $(1 - \rho(A))\lambda_j(A) \prod_i \Phi(A_i^j)$, is the marginal conditional likelihood of

$$\{\text{Partition } A \text{ in the } j\text{th way}, \mathbf{Y}(A) = \mathbf{y}(A)\} \text{ given } \{\mathbf{X}(A) = \mathbf{x}(A)\}.$$

Thus the conjugacy of the prior and the posterior updates for ρ , λ_j and $\text{OPT}(\mathcal{R}_{\mathbf{Y}}^A; \rho_{\mathbf{Y}}^A, \lambda_{\mathbf{Y}}^A, \alpha_{\mathbf{Y}}^A)$ follows from Bayes' Theorem and the posterior conjugacy of the standard optional Pólya tree prior (Theorem 3 in [40]). \square

Proof of Theorem 4. By Theorem 2.1 in [28], which follows directly from Schwartz's theorem (see [32] and [11, Theorem 4.4.2]), we just need to prove that the prior places positive probability mass in arbitrarily small Kullback-Leibler (K-L) neighborhoods of $f(\cdot|\cdot)$ w.r.t $f_{\mathbf{X}}$. Here a K-L neighborhood w.r.t $f_{\mathbf{X}}$ is defined to be the collection of conditional densities

$$K_{\epsilon}(f) = \left\{ h(\cdot|\cdot) : \int f(y|x) \log \frac{f(y|x)}{h(y|x)} f_{\mathbf{X}}(x) \mu(dx \times dy) < \epsilon \right\}$$

for some $\epsilon > 0$.

To prove this, we just need to show that any conditional density that satisfies the conditions given in the theorem can be approximated arbitrarily well in K-L distance by a piecewise constant conditional density of the sort that arises from the cond-OPT procedure. We first assume that $f(\cdot|\cdot)$ is continuous. Following the proof of Theorem 2, let $\delta(\epsilon)$ denote the modulus of continuity of $f(\cdot|\cdot)$. Let $\Omega_{\mathbf{X}} = \cup_{i=1}^K A_i$ be a reachable partition of $\Omega_{\mathbf{X}}$ such that the diameter of each partition block A_i is less than ϵ . Next, for each A_i , let $\Omega_{\mathbf{Y}} = \cup_{j=1}^N B_{ij}$ be a partition on $\Omega_{\mathbf{Y}}$ allowed under $\text{OPT}(\mathcal{R}_{\mathbf{Y}}; \rho_{\mathbf{Y}}^{A_i}, \lambda_{\mathbf{Y}}^{A_i}, \alpha_{\mathbf{Y}}^{A_i})$ such that the diameter of each B_{ij} is also less than ϵ . Let

$$g_{ij} = \sup_{x \in A_i, y \in B_{ij}} f(y|x) \quad \text{and} \quad g_i(y) = \sum_j g_{ij} I_{B_{ij}}(y).$$

Let $G_i = \int_{A_i \times \Omega_{\mathbf{Y}}} g_i(y) f_{\mathbf{X}}(x) \mu(dx \times dy)$. Then

$$0 \leq \sum_i G_i - 1 = \sum_i \int_{A_i \times \Omega_{\mathbf{Y}}} (g_i(y) - f(y|x)) f_{\mathbf{X}}(x) d\mu \leq \delta(2\epsilon) \mu_{\mathbf{Y}}(\Omega_{\mathbf{Y}}),$$

and so $\sum_i G_i \leq 1 + \delta(2\epsilon)\mu_{\mathbf{Y}}(\Omega_{\mathbf{Y}})$.

Now let $g(y|x) = \sum_i \left(g_i(y) / \int_{\Omega_{\mathbf{Y}}} g_i(\tilde{y}) \mu_{\mathbf{Y}}(d\tilde{y}) \right) I_{A_i}(x)$, which is a step function that can arise from the cond-OPT prior. Then

$$\begin{aligned}
0 &\leq \int f(y|x) \log(f(y|x)/g(y|x)) f_{\mathbf{X}}(x) d\mu \\
&= \sum_i \left(\int_{A_i \times \Omega_{\mathbf{Y}}} f(y|x) \log(f(y|x)/g_i(y)) f_{\mathbf{X}}(x) d\mu \right. \\
&\quad \left. + \int_{A_i \times \Omega_{\mathbf{Y}}} f(y|x) \log \left(\int_{\Omega_{\mathbf{Y}}} g_i(\tilde{y}) \mu_{\mathbf{Y}}(d\tilde{y}) \right) f_{\mathbf{X}}(x) d\mu \right) \\
&\leq \sum_i \log \left(\int_{\Omega_{\mathbf{Y}}} g_i(\tilde{y}) \mu_{\mathbf{Y}}(d\tilde{y}) \right) P(\mathbf{X} \in A_i) \\
&\leq \log \left(\sum_i \int_{\Omega_{\mathbf{Y}}} g_i(\tilde{y}) \mu_{\mathbf{Y}}(d\tilde{y}) P(\mathbf{X} \in A_i) \right) = \log(\sum_i G_i) \leq \delta(2\epsilon)\mu_{\mathbf{Y}}(\Omega_{\mathbf{Y}}),
\end{aligned}$$

which can be made arbitrarily close to 0 by choosing a small enough ϵ . Now if $f(\cdot|\cdot)$ is not continuous, then for any $\epsilon' > 0$, there exists a compact set $E \subset \Omega_{\mathbf{X}} \times \Omega_{\mathbf{Y}}$ such that $f(\cdot|\cdot)$ is uniformly continuous on E and $\mu(E^c) < \epsilon'$. Now let

$$g_{ij} = \left(\sup_{(x,y) \in E \cap (A_i \times B_{ij})} f(y|x) + \delta(\epsilon/2) \right) \vee \epsilon''$$

for some constant $\epsilon'' > 0$, while keeping the definitions of g_i , G_i and $g(y|x)$ in terms of g_{ij} unchanged. Let M be a finite upperbound of $f(\cdot|\cdot)$ and $f(\cdot, \cdot)$. We have

$$\begin{aligned}
\sum_i G_i - 1 &= \sum_i \int_{E \cap (A_i \times \Omega_{\mathbf{Y}})} (g_i(y) - f(y|x)) f_{\mathbf{X}}(x) d\mu \\
&\quad + \sum_i \int_{E^c \cap (A_i \times \Omega_{\mathbf{Y}})} (g_i(y) - f(y|x)) f_{\mathbf{X}}(x) d\mu.
\end{aligned}$$

Thus,

$$\sum_i G_i - 1 \geq \delta(\epsilon/2)\mu_{\mathbf{Y}}(\Omega_{\mathbf{Y}}) - (2M + \epsilon'')M\mu_{\mathbf{Y}}(\Omega_{\mathbf{Y}})\epsilon',$$

which is positive for small enough ϵ' . At the same time,

$$\sum_i G_i - 1 \leq (\delta(2\epsilon) + \epsilon'')\mu_{\mathbf{Y}}(\Omega_{\mathbf{Y}}) + (2M + \epsilon'')M\mu_{\mathbf{Y}}(\Omega_{\mathbf{Y}})\epsilon',$$

which can be made arbitrarily small by taking ϵ , ϵ' , and ϵ'' all $\downarrow 0$.

Now

$$\begin{aligned}
0 &\leq \int f(y|x) \log(f(y|x)/g(y|x)) f_{\mathbf{X}}(x) d\mu \\
&= \sum_i \left(\int_{A_i \times \Omega_{\mathbf{Y}}} f(y|x) \log(f(y|x)/g_i(y)) f_{\mathbf{X}}(x) d\mu \right. \\
&\quad \left. + \int_{A_i \times \Omega_{\mathbf{Y}}} f(y|x) \log \left(\int_{\Omega_{\mathbf{Y}}} g_i(\tilde{y}) \mu_{\mathbf{Y}}(d\tilde{y}) \right) f_{\mathbf{X}}(x) d\mu \right) \\
&= \sum_i \int_{E \cap (A_i \times \Omega_{\mathbf{Y}})} f(y|x) \log(f(y|x)/g_i(y)) f_{\mathbf{X}}(x) d\mu \\
&\quad + \sum_i \int_{E^c \cap (A_i \times \Omega_{\mathbf{Y}})} f(y|x) \log(f(y|x)/g_i(y)) f_{\mathbf{X}}(x) d\mu \\
&\quad + \sum_i \int_{A_i \times \Omega_{\mathbf{Y}}} f(y|x) \log \left(\int_{\Omega_{\mathbf{Y}}} g_i(\tilde{y}) \mu_{\mathbf{Y}}(d\tilde{y}) \right) f_{\mathbf{X}}(x) d\mu \\
&\leq 0 + M\epsilon' \log(M/\epsilon'') + \log(\sum_i G_i) \\
&\leq M\epsilon' \log(M/\epsilon'') + (\delta(2\epsilon) + \epsilon'') \mu_{\mathbf{Y}}(\Omega_{\mathbf{Y}}) + (2M + \epsilon'') M \mu_{\mathbf{Y}}(\Omega_{\mathbf{Y}}) \epsilon'.
\end{aligned}$$

The right hand side $\downarrow 0$ if we take $\epsilon \downarrow 0$ and $\epsilon' = \epsilon'' \downarrow 0$. This completes the proof. \square

References

- [1] BASHTANNYK, D. M. AND HYNDMAN, R. J. (2001). Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis* 36, 279–298.
- [2] CHIPMAN, H. A., GEORGE, E. I., AND MCCULLOCH, R. E. (1998). Bayesian CART model search. *Journal of the American Statistical Association* 93, 443, 935–948.
- [3] CHUNG, Y. AND DUNSON, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of The American Statistical Association* 104, 1646–1660.
- [4] DE IORIO, M., ROSNER, P., AND MACEachern, S. N. (2004). An anova model for dependent random measures. *Journal of The American Statistical Association* 99, 205–215.
- [5] DENISON, D. G. T., MALLICK, B. K., AND SMITH, A. F. M. (1998). A Bayesian CART algorithm. *Biometrika* 85, 2, 363–377. [MR1649118](#)
- [6] DUNSON, D. B. AND PARK, J.-H. (2008). Kernel stick-breaking processes. *Biometrika* 95, 307–323.
- [7] FAN, J., YAO, Q., AND TONG, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. *Biometrika* 83, 189–206.
- [8] FAN, J. AND YIM, T. H. (2004). A crossvalidation method for estimating conditional densities. *Biometrika* 91, 4 (Dec.), 819–834. <http://dx.doi.org/10.1093/biomet/91.4.819>.
- [9] FERGUSON, T. S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* 1, 209–230. [MR0350949 \(50 #3441\)](#)
- [10] GELFAND, A. E., KOTTAS, A., AND MACEachern, S. N. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association* 100, 1021–1035.
- [11] GHOSH, J. K. AND RAMAMOORTHY, R. V. (2003). *Bayesian Nonparametrics*. Springer Series in Statistics. Springer-Verlag, New York. [MR1992245 \(2004g:62004\)](#)
- [12] GRIFFIN, J. AND STEEL, M. (2006). Order-based dependent Dirichlet processes. *Journal of the American Statistical Association* 101, 179–194.
- [13] HALL, P., WOLFF, R. C., AND YAO, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association* 94, 445, 154–163. <http://eprints.qut.edu.au/5939/>.
- [14] HANSON, T. AND JOHNSON, W. O. (2002). Modeling regression error with a mixture of pólya trees. *Journal of the American Statistical Association* 97, 460.

- [15] HANSON, T. E. (2006). Inference for mixtures of finite pólya tree models. *Journal of the American Statistical Association* **101**, 476.
- [16] HYNDMAN, R. J. AND YAO, Q. (2002). Nonparametric estimation and symmetry tests for conditional density functions. *Nonpara. Statist* **14**, 259–278.
- [17] JARA, A. AND HANSON, T. E. (2011). A class of mixtures of dependent tail-free processes. *Biometrika* **98**, 3, 553–566.
- [18] LAVINE, M. (1992). Some aspects of Pólya tree distributions for statistical modelling. *Ann. Statist.* **20**, 3, 1222–1235. [MR1186248 \(93k:62006b\)](#)
- [19] LENK, P. J. (1988). The logistic normal distribution for Bayesian, nonparametric, predictive densities. *Journal of the American Statistical Association* **83**, 402, 509–516. <http://dx.doi.org/10.2307/2288870>.
- [20] LIU, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer.
- [21] MA, L. (2016). Adaptive shrinkage in Pólya tree type models. *Bayesian Analysis*. <http://projecteuclid.org/euclid.ba/1473276260>.
- [22] MA, L. AND WONG, W. H. (2011). Coupling optional Pólya trees and the two sample problem. *Journal of the American Statistical Association* **106**, 496, 1553–1565.
- [23] MACEACHERN, S. (1999). Dependent Dirichlet processes. In *Proceedings of the section on Bayesian Statistical Science*.
- [24] MALEK, M., TAGHIYAR, M. J., CHONG, L., FINAK, G., GOTTARDO, R., AND BRINKMAN, R. R. (2014). flowdensity: Reproducing manual gating of flow cytometry data by automated density-based cell population identification. *Bioinformatics*.
- [25] MAULDIN, R. D., SUDDERTH, W. D., AND WILLIAMS, S. (1992). Polya trees and random distributions. *Ann. Statist.*, 1203–1221.
- [26] MÜLLER, P., ERKANLI, A., AND WEST, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 1 (Mar.), 67–79. <http://dx.doi.org/10.1093/biomet/83.1.67>.
- [27] NORETS, A. (2011). Bayesian modeling of joint and conditional distributions. Tech. rep., Princeton University.
- [28] NORETS, A. AND PELENIS, J. (2011). Posterior consistency in conditional density estimation by co-variate dependent mixtures. Tech. rep., Princeton University.
- [29] PATI, D., DUNSON, D., AND TOKDAR, S. (2011). Posterior consistency in conditional distribution estimation. Tech. rep., Duke University Department of Statistical Science.
- [30] RODRÍGUEZ, A. AND DUNSON, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis* **6**, 1, 145–178.
- [31] ROEDER, K. AND WASSERMAN, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association* **92**, 894–902.
- [32] SCHWARTZ, L. (1965). On Bayes procedures. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **4**, 10–26. [MR0184378 \(32 #1851\)](#)
- [33] SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.
- [34] SHEN, W. AND GHOSAL, S. (2016). Adaptive bayesian density regression for high-dimensional data. *Bernoulli* **22**, 1 (02), 396–420. <http://dx.doi.org/10.3150/14-BEJ663>.
- [35] TADDY, M. A. AND KOTTAS, A. (2010). A Bayesian nonparametric approach to inference for quantile regression. *Journal of Business & Economic Statistics* **28**, 3, 357–369.
- [36] TOKDAR, S. T. AND GHOSH, J. K. (2007). Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference* **137**, 1 (Jan.), 34–42. <http://dx.doi.org/10.1016/j.jspi.2005.09.005>.
- [37] TOKDAR, S. T., ZHU, Y. M., AND GHOSH, J. K. (2010). Bayesian density regression with logistic gaussian process and subspace projection. *Bayesian analysis* **5**, 2, 319–344.
- [38] TRIPPA, L., MLLER, P., AND JOHNSON, W. (2011). The multivariate beta process and an extension of the polya tree model. *Biometrika* **98**, 1, 17–34.
- [39] WALKER, S. G. AND MALLICK, B. K. (1997). Hierarchical generalized linear models and frailty models with Bayesian nonparametric mixing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**, 4, 845–860. <http://dx.doi.org/10.1111/1467-9868.00101>.
- [40] WONG, W. H. AND MA, L. (2010). Optional Pólya tree and Bayesian inference. *Annals of Statistics* **38**, 3, 1433–1459.